Automated Research Workflows for Accelerated Discovery

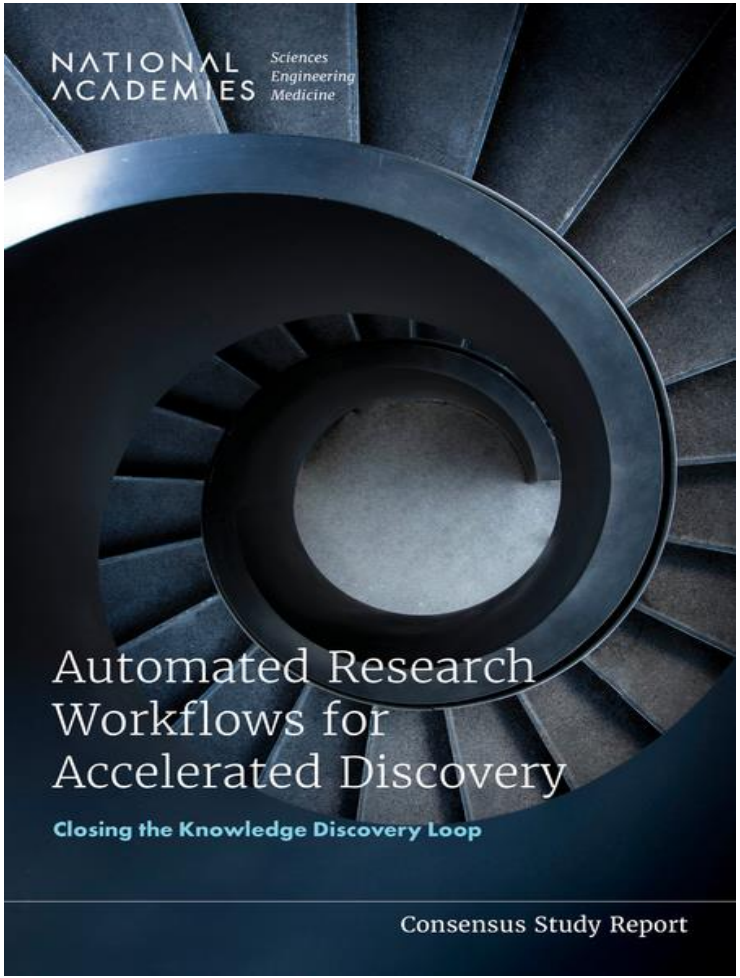
Briefing for the Department of Energy's Biological and Environmental Research Advisory Committee

October 14, 2022

Tapio Schneider, Committee Member

Theodore Y. Wu Professor of Environmental Science and Engineering

California Institute of Technology



Outline

- Background and context for the report
- Overview of Automated Research Workflows (ARWs)
- Examples and use cases
- Findings and recommendations
- Future topics and questions

National Academies of Sciences, Engineering, and Medicine

- Private, non-profit, self-selecting membership organizations
- Congressional charter to advise the Federal government
- 6 major divisions, boards/standing committees, ad hoc committees
- Consensus studies, convening activities, operating programs

Automated Research Workflows project

- Consensus study organized under NASEM's Board on Research Data and Information
- Sponsored by Schmidt Futures
- Study goals: (1) Examine current efforts to develop advanced and automated workflows for scientific research, (2) Identify promising research approaches to promote progress by the community in creating and using more productive research workflow systems.
- Study process: (1) Committee meetings, (2) March 2020 workshop, (3) Literature review
- Final report released May 2022; includes recommendations for strategic actions to be taken by multiple stakeholders to seize the perceived opportunities.

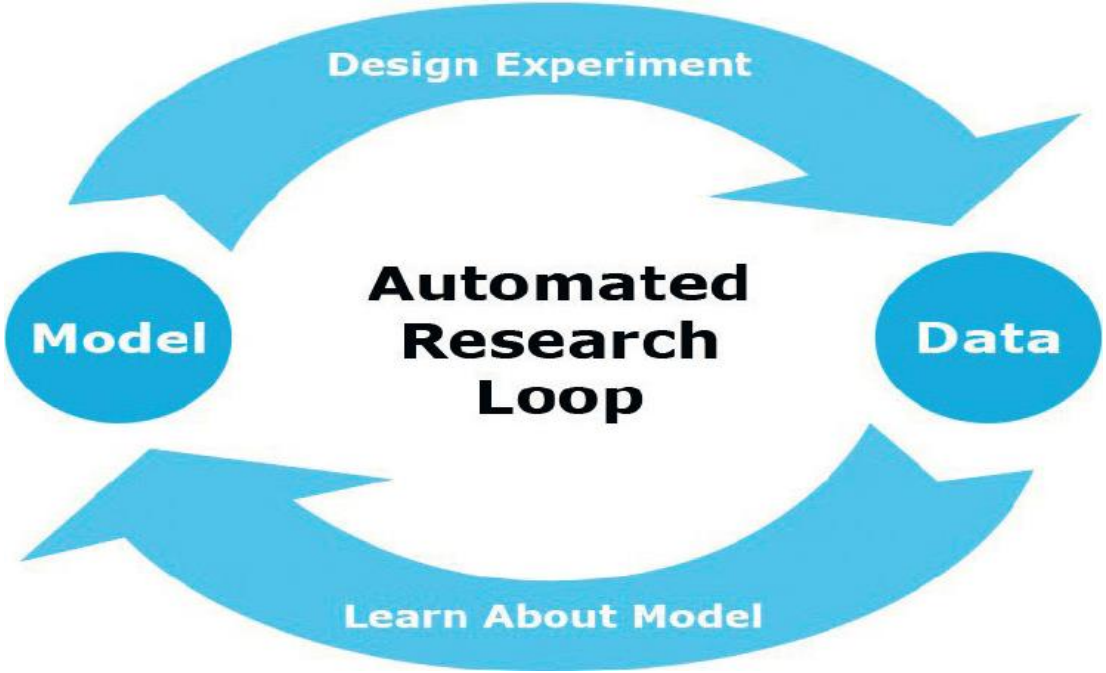
Study committee members and NASEM staff

- Daniel Atkins, University of Michigan
- Ilkay Altintas, San Diego Supercomputer Center
- Shreyas Cholia, Lawrence Berkeley National Laboratory
- Mercé Crosas, Secretary of Open Government, Government of Catalunya
- Alfred Hero, University of Michigan
- Rebecca Lawrence, F1000 Research Limited, London
- Bradley Malin, Vanderbilt University
- Lara Mangravite, Sage Bionetworks
- Tapio Schneider, California Institute of Technology
- Tom Arrison, National Academies Study Director
- Emi Kameyama, National Academies Program Officer

Overview of Automated Research Workflows

- ARWs are scientific research processes emerging across a variety of disciplines and fields
- ARWs integrate computation, laboratory automation, and tools from AI in the performance of tasks that make up the research process, such as designing experiments-observations-simulations, collecting and analyzing data, and learning from the results to inform further experiments-observations-simulations
- Specific tools and resources vary by field, but the common goal is to accelerate scientific knowledge generation while achieving greater control and reproducibility

Knowledge discovery loop



Automated research workflows can automate and close the loop of scientific discovery. On one side of the loop, artificial intelligence (AI) and machine learning (ML) algorithms harness the experimental or observational data to learn about a model; on the other side of the loop, AI and ML are used to generate the study design for the next data collection. The loop goes on iteratively.

ARWs are a next step in the research computing revolution

- Stage One: Computers as a programmed calculator for science and engineering calculations.
- Stage Two: Centralized, center-based computers for general purpose computing including science and engineering.
- Stage Three: Specialized research supercomputers plus network.
- Stage Four: Distributed, ubiquitous computing and the rise of the all-digital world.
- Stage Five: Cyberinfrastructure ecologies of computing, data, information, acquisition, activation and distance-independent collaboration. (collaboratory, e-science)
- Stage Six: Stage Five augmented by artificial intelligence and automation/robotics to make the workflow of research more efficient and productive. (accelerate discovery)

Why now?

- Growing wide-spread awareness of the transformative potential of AI as a research tool
- Significant progress in open science (including open data and publications) that can now be leveraged in workflow systems to help make data and other research objects more findable, accessible, interoperable, and reusable (FAIR).
- The growing use of contemporary digital lab books (e.g. Jupiter), and workflow system, that can be on-ramps to ARWs.
- Increasing challenges to reliable and reproducible research that workflow systems can address
- Potential for the acceleration of scientific discovery to help meet the pressing grand challenges in our world

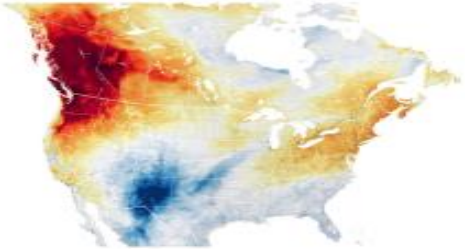
Examples of ARWs in action

- Material Science - Cut the time required for synthesis and testing of materials from 9 months to 5 days
- Particle Physics - Allow experiments to achieve a given sensitivity with ½ the data
- Drug Discovery - Identify 57 percent of active compounds performing 2.5% of possible experiments, compared with 20% identified with traditional approach
- Astronomy - Automate telescope target selection so that observations are optimally informative given constraints and scientific objectives.
- Digital Humanities - Compile information from enormous volumes of words across multitudes of languages over centuries to see patterns in how ideas have spread and changed over time, and to understand the development of human thought.

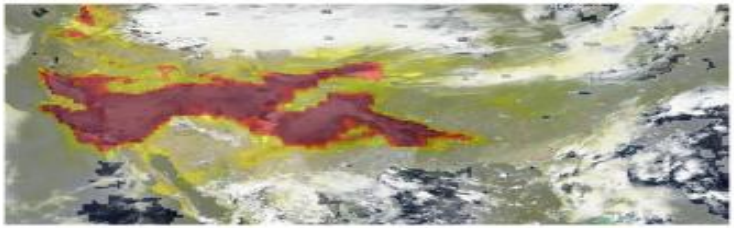
Use case: Climate modeling

Losses from weather and climate disasters in the U.S. have increased from $30B/yr in 1990s to $150B/yr now

Erftstadt-Blessem, Germany, July 15, 2021 (AP)



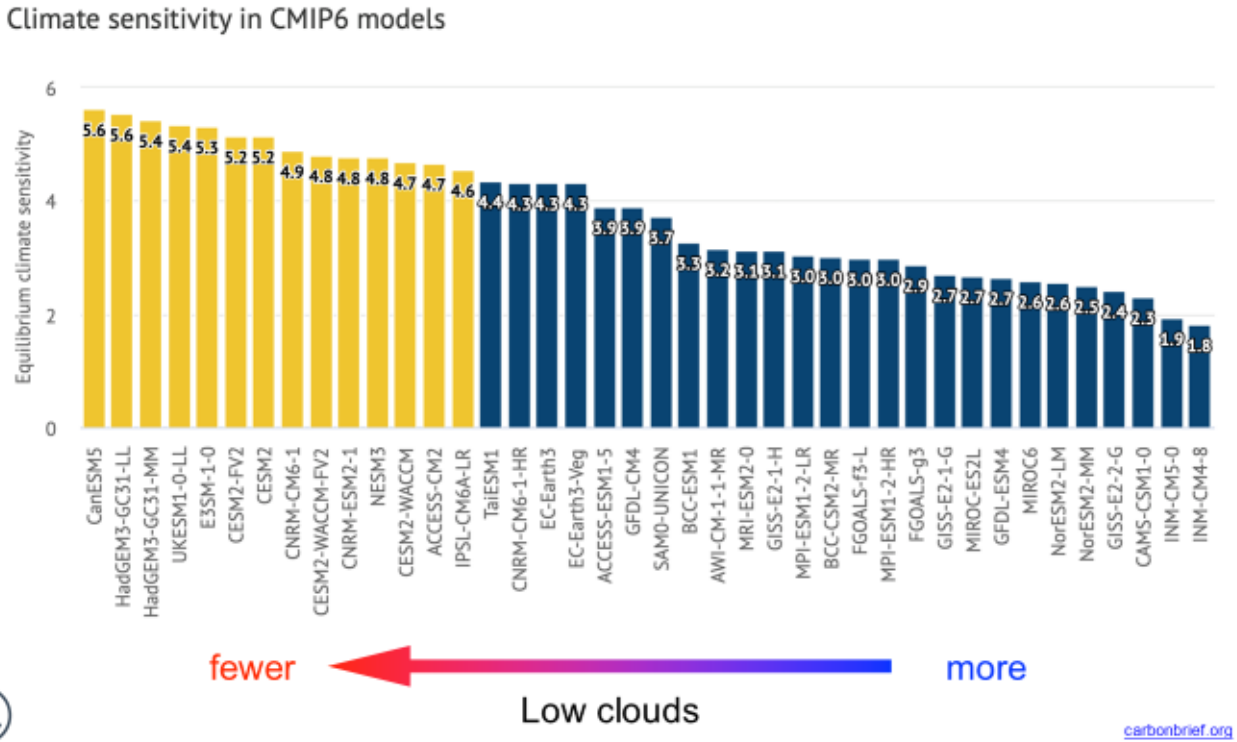
June 2021 heat wave (NASA)



Smoke from California wildfires Sept. 7, 2020 (OMPS, Suomi NPP)

NOAA National Centers for Environmental Information. U.S. Billion-Dollar Weather and Climate Disasters, 2022

Adaptation to climate change is unavoidable, but what to adapt to is uncertain, especially tail risks



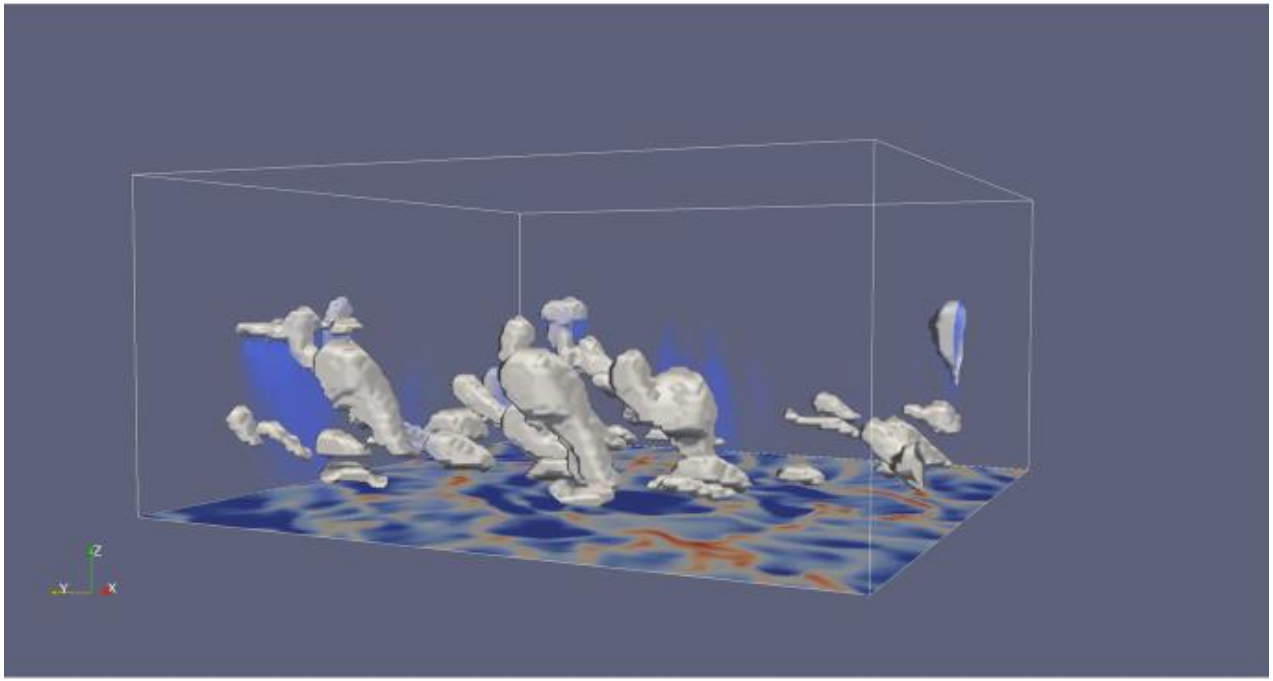
Clouds and other small-scale processes dominate uncertainties in climate projections

http://eoimages.gsfc.nasa.gov

Stratocumulus: colder

Cumulus: warmer

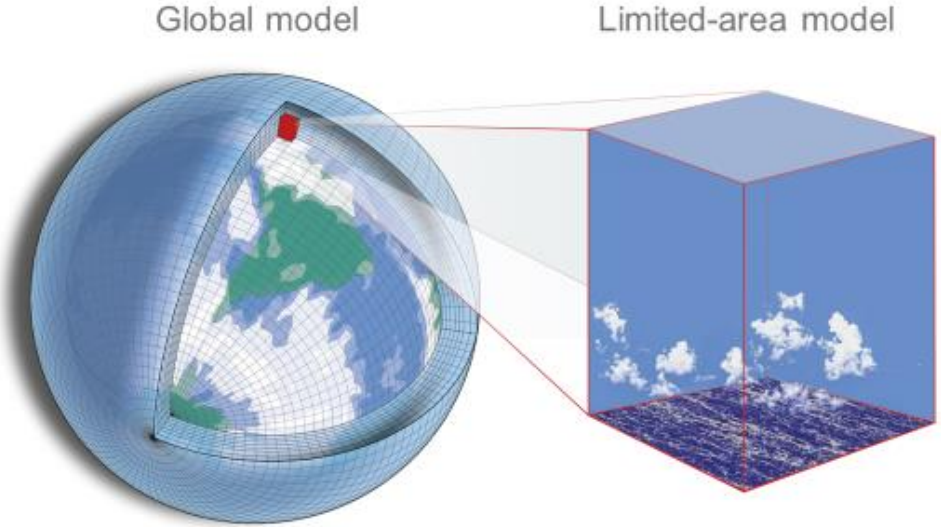
While we cannot resolve clouds globally—it would require 100 billion times current compute—we can resolve them locally



Large-eddy simulation of tropical cumulus

Simulation with PyCLES (Pressel et al. 2015)

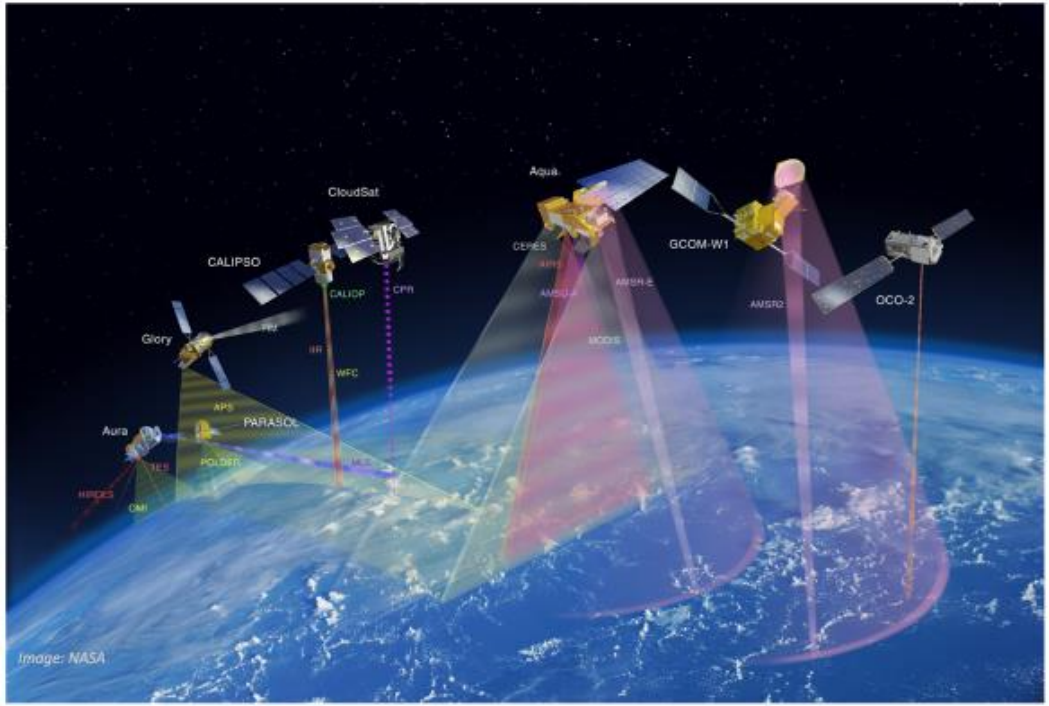
Such limited area models can be nested in a global model and can, in turn, inform the global model



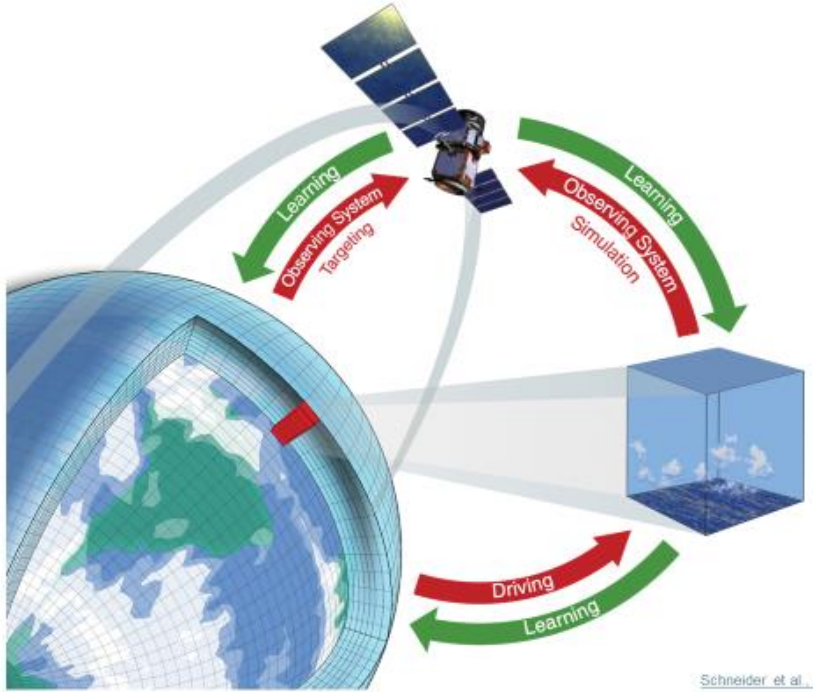
Thousands or tens of thousands of high-resolution simulations can be embedded in a global model, and the global model can learn from them

Schneider et al., Nature Climate Change 2017

We have global climate measurements, whose potential to improve models has not been tapped

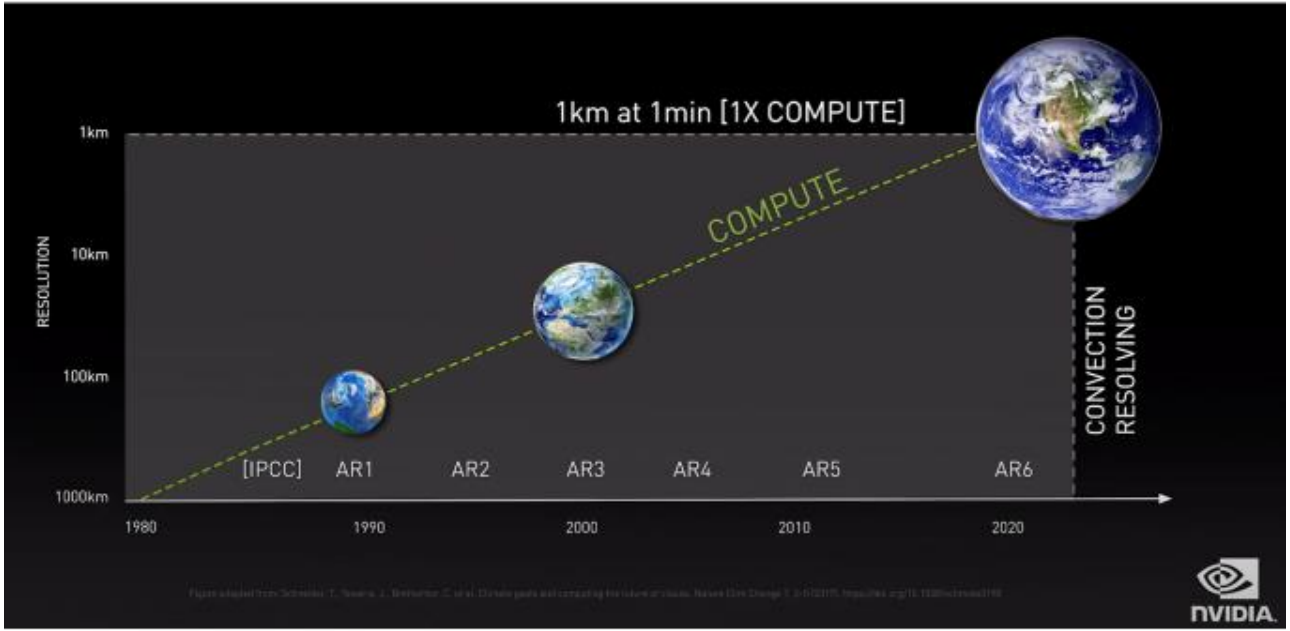


Progress can be accelerated by automating the workflow of designing high-resolution simulations and learning from them



Schneider et al., *Geophys. Res. Lett.*, 2017

Brute-force computing alone cannot solve the problem



Schneider et al. *Nature Climate Change*, 2017

To realize such an automated workflow, we need new AI-based algorithms

- **Experimental design:** At which location will the next high-resolution simulation to be maximally informative?
  - Various approaches available (e.g., Dunbar et al. 2022 in the climate context)
- **Learning from data:** How can we learn from disparate and noisy data (including observations), within physics-based models?
  - Bayesian learning approaches suggest themselves, but are computationally expensive
  - AI-accelerated Bayesian learning achieves 1000x speedup, making it feasible for climate models and other computationally expensive models (Cleary et al. 2021)

These algorithms transcend individual fields; they are beginning to be developed but one day may be as ubiquitous as least squares is now

Four major report findings

1. ARWs have the potential to integrate computation, data acquisition and storage, laboratory automation, and tools from AI in the performance of tasks that make up the research process.
2. ARWs can potentially accelerate discovery as well as foster reproducibility, replicability, and rigor; facilitate collaboration across disciplines and nations
3. Advancement in the creation and application of ARWs requires investment and changes in the research enterprise: (a) Sustainable funding, (b) Appropriate education, retraining, (c) Reporting and sharing methods and results, (d) Changes in rewards and incentives, (e) Multi-role, multi disciplinary collaboration.
4. Legal and policy issues must be addressed and some on an international basis especially for personal and medical data.

Report recommendations

1. ARW design principles should (a) Facilitate openness, reproducibility, and transparency, (b) Incorporate principles of responsible AI, (c) Prioritize reuse and sustainability, (d) Be driven and controlled by the research community.
2. Further progress on openness, sustainability and sharing of infrastructure, instruments, code, and data is required.
3. Research funders, performers, their institutions and professional societies should cooperate in supporting the education and training required for creating and using ARWs.

Report recommendations, continued

4. An enhanced culture of sharing with incentives to do so is critical to the creation of ARW-based research.
5. Preservation of privacy must be robustly addressed in ARW world.

What next?

- Recent research and policy initiatives, such as the Artificial Intelligence Initiative and OSTP Open Access Guidance will support progress toward implementation of ARWs across more disciplines
- Schmidt Futures is launching a new program aimed at training and educating researchers in the areas of knowledge/expertise needed to develop and utilize ARWs
- More needs to be done to address culture/incentive issues within institutions and disciplines
- We continue to look for opportunities to discuss the report

Thank you and questions

For information about the study, printed copies, etc.:

Tom Arrison

Director, Board on Research Data and Information

National Academies of Sciences, Engineering, and Medicine

Email: tarrison@nas.edu